Research Statement

Yin Lin, Ph.D. Candidate (irenelin@umich.edu)

My primary research interests are centered on responsible data management, aiming to mitigate the harmful outputs of AI systems. Big data are extensively used to train AI systems and to develop algorithms for various decision-making tasks. However, these data-driven systems and algorithms are only as reliable as the data they are built on, which often contain human biases embedded in the pipelines or reflect historical biases inherited from the collected datasets [1]. Without proper utilization and investigation, this "bias in, bias out" issue prevents AI-based systems from producing equitable outcomes, leading to undesirable biases, particularly against minorities frequently underrepresented in datasets and overlooked in data management processes.

My research inspects the data pipeline of AI systems to enhance fairness and transparency in the use of big data. From a data researcher's perspective, my work revolves around two primary goals: achieving **intrinsic fairness** within the data pipeline and mitigating **extrinsic unfairness** inherited from collected datasets. Ultimately, my research aims to (1) provide reliable data management platforms that enable data scientists and machine learning engineers to prepare their data for downstream usage and (2) perform analytical evaluations and remediations of their datasets for fairness purposes.

1. Recent and Ongoing Research

My thesis specifically addresses intersectionality in responsible data management. Instead of focusing on a small set of predefined protected groups based on a single dimension (e.g., race or gender), my work examines the intersectional effects of multiple protected attributes. Intersectionality is crucial in fairness, which recognizes and addresses the impact of overlapping forms of oppression (e.g. racism and sexism). For example, we cannot oversimplify the obstacles faced by Black female workers to the sum of the obstacles respectively faced by Black workers and female workers considered separately [2]. Considering intersectionality in responsible data management presents unique challenges in both modeling—identifying occurrences of unfairness given the complexity of identities—and computational complexity, given the power set of subgroups in the intersectional space to explore.

To address these challenges, I have developed a suite of techniques that involve both modeling unfairness and creating efficient algorithms to tackle these issues. For example, inspired by real-world examples where news or reporting could *cherry-pick* generalized conclusions that do not accurately reflect the experiences of all subgroups, my research proposed a framework to evaluate, capture, and revise such misleading conclusions. Additionally, I studied *row-level data lineage* in data science pipelines with non-relational operators and user-defined functions (e.g., machine learning and data analytical pipelines). Given an erroneous or individual record at the pipeline's output, my work provided an efficient platform to trace its lineage through each pipeline operator, identifying the source of the specified record. For the training set pre-processing analysis, my research specifically considers representation bias, where certain populations in datasets are underrepresented. I formally illustrated the connection between representation bias in training data and outcome divergence among subgroups in machine learning prediction results, proposing mitigation strategies based on *dataset remedy* and *coverage analysis*. Furthermore, my research explored *AI-driven* approaches for data manipulation to enhance features and improve model performance.

a. Cherry-picking generalization

Creating generalizations from detailed data is crucial for understanding extensive datasets. However, poorly constructed generalizations can mislead even when technically supported. Evaluating these generalizations is vital in data analytics to avoid misinterpretation. For instance, despite the popularity of Sex and the City, it appears less favored than Gossip Girl according to IMDb ratings (7.1 vs. 7.4). This discrepancy arises because Sex and the City appeals primarily to women, who rate it higher (8) compared to Gossip Girl (7.7). However, male raters give Sex and the City significantly lower ratings, affecting the overall score. When there is a large difference, such as in this case, it is essential to include these details in the produced generalization.

In [3][4], I built a scoring framework for detecting and explaining faulty generalizations by refining aggregate queries along a set of partition attributes. My approach evaluated generalizations by considering the sizes of subgroups as weights for support. By utilizing refinement queries, it assessed how subgroup partitions supported or opposed the aggregate statement. The statement's score indicated the accuracy of the aggregation in representing the underlying data. I refined aggregates using typical OLAP operations, such as *slicing* and *drill-down*, resulting in a powerset of the subgroups to explore. My proposed algorithm started by constructing a hierarchy and designing efficient algorithms to traverse all potential partitions. This method demonstrated superior efficiency compared to naive and database CUBE methods. I validated our framework's capabilities by analyzing real-world data statements and providing an interactive demo.

b. Row-level data lineage for data science pipelines

Data lineage is important in data processing, describing how output data items are derived from input data through a series of transformations, with practical applications in data debugging, GDPR compliance, and data integration. Prior work includes eager lineage tracking and lazy lineage inference. Eager tracking proactively tracks lineage during query execution, enabling customized tracking and efficient lineage querying, but it is often deeply integrated with a specific DBMS and hard to apply to other platforms. Lazy inference generates additional queries to compute lineage, making it easily applicable to any data platform, but the lineage query is usually less efficient. Both approaches have limited support for *non-relational operators, user-defined functions (UDFs)*, and *correlated nested queries*, which are common in real-world data science pipelines.

I proposed PredTrace [5], which leverages the predicate pushdown technique to achieve easy adaptation, low runtime overhead, efficient lineage querying, and high pipeline coverage. PredTrace utilizes a search-verification-based component [6] that pushes a row-selection predicate that describes the target output row, down to input tables and queries the lineage by executing the pushed predicate. To obtain precise lineage, PredTrace may require saving intermediate results during pipeline execution. I devised techniques to reduce the size of materialized intermediate results. When saving intermediate results is not viable, PredTrace can still infer lineage but may return a superset. My approach also includes an algorithm to generate the most useful lineage supersets, ensuring they are not excessively large. Compared to prior work, PredTrace achieves higher coverage on all TPC-H queries as well as 70 sampled real-world data processing pipelines where UDFs are widely used. It can infer lineage in seconds, outperforming state-of-the-art lineage baselines.

c. Exploratory pre-training analysis and remedy

In data-driven applications, it is critical to use appropriate datasets for analysis and training purposes. However, due to factors such as historical discrimination or data scientists' limited control over the sampling processes, representation bias often occurs in the collected datasets. This results in the data failing to adequately represent the population, particularly minorities. Therefore, performing pre-training analysis to ensure comprehensive representation across different groups over the intersection of multiple attributes is an important step to mitigate potential unfairness within these groups.

In [7], my research addressed the issue of inadequate data collection in databases with multiple relations. The goal is to identify the maximum uncovered groups in the intersectional space of protected attributes, where their coverage falls below a specified threshold. I tackled this challenge using efficient algorithms and index schemas, as no polynomial-time solutions exist. Additionally, accessing coverage for multi-table databases requires complex joins and predicate combinations. My approach involved creating a compact, parallel index for efficient counting, and leveraging monotonicity and a priority-based heuristic for identifying maximum uncovered groups. I also incorporated sampling-based techniques to reduce time overhead in coverage analysis further.

Furthermore, in [8], my research uncovered a significant correlation between representation bias in training data and model fairness, particularly in achieving similar performance across protected groups. Unlike traditional approaches that focus on predefined groups, my work delves into intersectional fairness, targeting arbitrary subgroups in the intersection of protected attributes. Through empirical and theoretical analyses, I showed how performance divergence arises when positive and negative instances are not proportionately represented within these intersectional regions, regardless of the learning algorithm used downstream. For instance, consider the COMPAS algorithm, where an overabundance of positive instances for African American males in the training set leads to a higher false positive rate (FPR) for this subgroup. I developed an algorithm to detect such skewed data collections in datasets, underlining the need to address biased data collection through pre-processing to achieve unbiased class distributions within intersectional regions. This approach not only proves effective but also demonstrates efficiency compared to various in-processing and pre-processing bias mitigation methods.

d. LLM for feature enhancement

Raw data collected through data integration is seldom suitable for direct use in machine learning or data analytics, as it typically requires extensive data wrangling to construct high-quality features. Leveraging large language models (LLMs) can facilitate the creation of new features based on contextual information and open-world knowledge. A key concern is exploring the efficient utilization of LLMs for data-wrangling tasks, considering the time and financial costs involved in interacting with the LLMs.

In [9], I proposed SMARTFEAT, an efficient automated feature engineering (AFE) tool to help data users construct useful features through *feature-level interaction* with LLMs. My method incorporates an intelligent *operator selector* that identifies a subset of operators, avoiding the exhaustive expansion-selection process as in traditional AFE tools. Additionally, I addressed the limitations of performing data tasks through *row-level interactions* with LLMs, which can cause significant delays and costs due to excessive API calls. I introduced a *function generator* that facilitates efficient data transformations, such as dataframe built-in methods or lambda functions, ensuring SMARTFEAT can generate new features for large datasets. Recognizing the inherent uncertainty in the automatic execution of AI-generated code, SMARTFEAT includes mechanisms to avoid generation errors and a feature selection component to enhance the quality of generated features. In evaluations on real-world datasets across various fields, SMARTFEAT demonstrated superior effectiveness, efficiency, and explainability compared to other AFE tools.

2. Motivation for Future Research

My current research paves the way for various promising future research directions. These avenues offer exciting opportunities to further advance the ethical and responsible applications of AI techniques. Some potential directions I plan to continue exploring include, but are not limited to, the following perspectives:

a. Responsible management of non-tabular data

Currently, our focus has been primarily on responsible data management for structured tabular data. However, with the expanding applications of AI techniques in computer vision, natural language processing, and graph-based learning, there is a growing need for responsible data management of non-tabular data such as images, text, and graphs. For instance, [10] evaluates three commercial face recognition systems and identifies that darker-skinned females consistently experience systematic unfairness due to the phenotypic and demographic representation in face datasets, as well as the quality of the collected images. As surveyed in [1], I have discussed techniques for identifying and resolving representation bias in non-tabular data. However, mitigating unfairness in these tasks remains far from satisfactory due to the complexity of these data types.

b. Responsible AI considerations for using generative AI

With the rapid development of generative AI applications, evaluating and mitigating the responsible concerns of these systems has become a critical topic among AI providers and users. For example, the content generated by language models has the potential to further marginalize groups that experience exclusion from social life [11]. Recent research advancements have developed frameworks and benchmarks to measure biases against such populations in language generation and approaches like adding guardrail models to the system output, producing clean data for pretraining, and engaging human beings to avoid such biases through reinforcement learning based on human feedback (RLHF) have been widely discussed. Once again, the data used to train these sophisticated models appears to be taking center stage. I am thrilled by the potential of this area, as generative AI opens up many opportunities while raising various concerns that need careful attention.

c. Provenance for generative AI

Data provenance addresses the origins and transformations of data, answering questions about how an output tuple is generated and why it exists in SQL queries or data pipeline outputs. While LLMs like ChatGPT and new Bing are now extensively used for generating textual content, they face a significant challenge known as hallucination—when an LLM generates text not supported by the input. To improve the reliability of such tools, incorporating provenance tracking for generative AI models is important. This involves understanding the sources and processes that produce the outputs and identifying whether specific user data was utilized. Therefore, I intend to investigate data provenance in generative AI, aiming to promote accountability and enhance trustworthiness.

References

[1] Shahbazi, N., Lin, Y., Asudeh, A., & Jagadish, H. V. (2023). Representation bias in data: a survey on identification and resolution techniques. ACM Computing Surveys, 55(13s), 1-39.

[2] Crenshaw, K. (2013). Demarginalizing the intersection of race and sex: A black feminist critique of antidiscrimination doctrine, feminist theory and antiracist politics. In Feminist legal theories (pp. 23-51). Routledge.

[3] Lin, Y., Youngmann, B., Moskovitch, Y., Jagadish, H. V., & Milo, T. (2021). On detecting cherry-picked generalizations. Proceedings of the VLDB Endowment (PVLDB), 15(1), 59-71.

[4] Lin, Y., Youngmann, B., Moskovitch, Y., Jagadish, H. V., & Milo, T. (2022). OREO: detection of cherry-picked generalizations. Proceedings of the VLDB Endowment (PVLDB), 15(12), 3570-3573.

[5] Lin, Y., & Yan, C. (in progress). Ongoing work.

[6] Yan, C., Lin, Y., & He, Y. (2023). Predicate pushdown for data science pipelines. Proceedings of the ACM on Management of Data (SIGMOD), 1(2), 1-28.

[7] Lin, Y., Guan, Y., Asudeh, A., & Jagadish, H. V. (2020). Identifying insufficient data coverage in databases with multiple relations. Proceedings of the VLDB Endowment (PVLDB), 13(11), 2229-2242.

[8] Lin, Y., Gupta, S., & Jagadish, H. V. (2024). Mitigating Subgroup Unfairness in Machine Learning Classifiers: A Data-Driven Approach. IEEE International Conference on Data Engineering (ICDE).

[9] Lin, Y., Ding, B., Jagadish, H. V., & Zhou, J. (2024). SMARTFEAT: Efficient Feature Construction through Feature-Level Foundation Model Interactions. Conference on Innovative Data Systems Research (CIDR).

[10] Buolamwini, J., & Gebru, T. (2018, January). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on fairness, accountability and transparency (pp. 77-91). PMLR.

[11] Ovalle, A., Goyal, P., Dhamala, J., Jaggers, Z., Chang, K. W., Galstyan, A., ... & Gupta, R. (2023). “I’m fully who I am”: Towards Centering Transgender and Non-Binary Voices to Measure Biases in Open Language Generation. In Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT) (pp. 1246-1266).